US006389436B1

(12) **United States Patent**
Chakrabarti et al.

(10) **Patent No.: US 6,389,436 B1**
(45) **Date of Patent: *May 14, 2002**

(54) **ENHANCED HYPERTEXT CATEGORIZATION USING HYPERLINKS**

(75) Inventors: **Soumen Chakrabarti**, San Jose; **Byron Edward Dom**, Los Gatos; **Piotr Indyk**, Stanford, all of CA (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: This patent issued on a continued prosecution application filed under 37 CFR 1.53(d), and is subject to the twenty year patent term provisions of 35 U.S.C. 154(a)(2).

Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **08/990,292**

(22) Filed: **Dec. 15, 1997**

(51) **Int. Cl.**[7] **G06F 15/00**
(52) **U.S. Cl.** **707/513**; 737/3; 737/4; 737/6
(58) **Field of Search** 707/513, 1–7; 706/20

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,274,744	A *	12/1993	Yu et al.	706/20
5,325,298	A	6/1994	Gallant	707/5
5,414,781	A *	5/1995	Spitz et al.	382/296
5,418,948	A *	5/1995	Turtle	707/4
5,463,773	A *	10/1995	Sakakibara et al.	707/102
5,488,725	A *	1/1996	Turtle et al.	707/5
5,594,897	A *	1/1997	Goffman	707/102
5,617,488	A *	4/1997	Hong et al.	382/229
5,642,502	A *	6/1997	Driscoll	707/5
5,675,710	A *	10/1997	Lewis	706/2
5,675,819	A *	10/1997	Schuetze	707/3
5,694,559	A *	12/1997	Hobson et al.	707/3
5,794,236	A *	8/1998	Mehrle	707/5
5,832,470	A *	11/1998	Morita et al.	707/1
5,835,905	A *	11/1998	Pirolli et al.	707/3
5,873,056	A *	2/1999	Liddy et al.	707/3
5,873,107	A *	2/1999	Borovoy et al.	707/501
5,895,470	A *	4/1999	Pirolli et al.	707/102
5,930,788	A *	7/1999	Wical	707/5
5,950,187	A *	9/1999	Tsuda	707/3
6,026,399	A *	2/2000	Kohavi et al.	707/6

OTHER PUBLICATIONS

Lu et al., Stereo Image Matching Based on Probability Relaxation, TENCON '97, IEEE Region 10 Annual Conference, Dec. 1997, vol. 1, pp. 315–318.*

Kato et al., Multicast Markov Random Field Models for Parallel Image Classification, Computer Vision, May 1993, pp. 253–257.*

S. Li, et al., "Relaxation Labeling of Markov Random Fields," IEEE, pp. 488–492, 1994.

J. Savoy, An Extended Vector—Processing Scheme For Searching Information in Hypertext Systems, Information Processing & Management, vol. 32, No. 2, pp. 155–170, 1996.

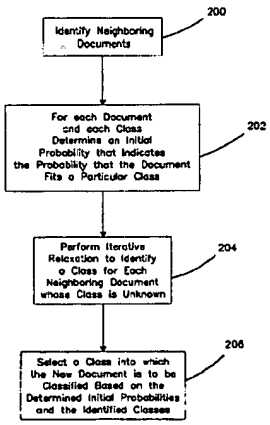
(List continued on next page.)

Primary Examiner—Stephen S. Hong
Assistant Examiner—Cong-Lac Huynh
(74) *Attorney, Agent, or Firm*—Altera Law Group, LLC

(57) **ABSTRACT**

A method, apparatus, and article of manufacture for a computer implemented hypertext classifier. A new document containing citations to and from other documents is classified. Initially, documents within a neighborhood of the new document are identified. For each document and each class, an initial probability is determined that indicates the probability that the document fits a particular class. Next, iterative relaxation is performed to identify a class for each document using the initial probabilities. A class is selected into which the new document is to be classified based on the initial probabilities and identified classes.

51 Claims, 15 Drawing Sheets



US-PAT-NO: 6389436

IBM

DOCUMENT-IDENTIFIER: US 6389436 B1

TITLE: Enhanced hypertext categorization using hyperlinks

DATE-ISSUED: May 14, 2002

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Chakrabarti; Soumen	San Jose	CA	N/A	N/A
Dom; Byron Edward	Los Gatos	CA	N/A	N/A
Indyk; Piotr	Stanford	CA	N/A	N/A

US-CL-CURRENT: 715/513, 707/3 , 707/4 , 707/6

ABSTRACT:

A method, apparatus, and article of manufacture for a computer implemented hypertext classifier. A new document containing citations to and from other documents is classified. Initially, documents within a neighborhood of the new document are identified. For each document and each class, an initial probability is determined that indicates the probability that the document fits a particular class. Next, iterative relaxation is performed to identify a class for each document using the initial probabilities. A class is selected into which the new document is to be classified based on the initial probabilities and identified classes.

51 Claims, 19 Drawing figures

Exemplary Claim Number: 1

Number of Drawing Sheets: 15

---------- KWIC ---------

Detailed Description Text - DETX (71):

A similar problem appears in inductive logic programming, as discussed in "Lavrac and Dzeroski". Suppose a relational table, or several tables, are given containing information about patients with hypertension. Apart from local attributes of a person, such as age and weight, there can be other attributes, such as father and mother, whose records are also in the database. Assuming there is a relation between lineage and hypertension, the latter fields are used for inducing rules to classify patients as having high or low risk of heart attack. To do this, each record is augmented with fields from records of the parents. Rather than using **raw data** from the related records, the raw attribute values can be pre-processed into a synthesized feature. One example is whether the parent was classified as high or low risk. This is called feature engineering.

Detailed Description Text - DETX (147):

6.1 Bridges "Co-citation" is a well-studied phenomenon in implicitly linked corpora, such as academic papers, as discussed in G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, 1983, which is

incorporated by reference herein. Documents that cite or are cited by many common documents may be regarded as similar, much as documents sharing many terms are adjudged similar. Citation-based similarity, or a combination of citation and term-based similarity, can then be used to perform **unsupervised clustering** of documents, discussed in R. Weiss, B. Velez, M. A. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. K. Gifford, "HyPursuit: A Hierarchical Network search Engine that Exploits Content-link Hypertext Clustering", Proc. of the Seventh ACM Conference on Hypertext, Washington, D.C., March 1996, which is incorporated by reference herein.